VC dimension
and
Model Selection

Overview

- PAC model: review
- VC dimension:
  - Definition
  - Examples
  - Sample:
    - Lower bound
    - Upper bound !!!
- Model Selection

PAC model: Setting

- A distribution: D (unknown)
- Target function: c_t from C
  - $c_t : X \rightarrow \{0,1\}$
- Hypothesis: h from H
  - $h: X \rightarrow \{0,1\}$
- Error probability:
  - $error(h) = Prob_D[h(x) \neq c_t(x)]$
- Oracle: $EX(c_t,D)$

PAC model: Definition

- C and H are concept classes over X.
- C is PAC learnable by H if
- There Exist an Algorithm A such that:
  - For any distribution D over X and c_t in C
  - for every input ε and δ:
  - outputs a hypothesis h in H,
    - while having access to $EX(c_t,D)$
  - with probability $1-\delta$ we have $error(h) < \varepsilon$
- Complexities: sample, running time

PAC model – last week

- For a finite hypothesis class H, sample size *m*:
  - Realizable case:

    $$m > (1/\varepsilon) \ln (|H|/\delta)$$

  - Non-realizable

    $$m > (2/\varepsilon^2) \ln (2|H|/\delta)$$

- Impossibility results:
  - $m > (1/2) \log |H|$
  - $m > (1/4\varepsilon) \ln (1/\delta)$

VC dimension: motivation

- Infinite hypothesis class
  - Threshold
  - Rectangles
  - TODAY: general
    - VC dimension
    - Applies both to realizable and non-realizable.

VC dimension: definition

- Notation: C – concept class; S - sample
- Projection: $\Pi_C(S) = \{c \cap S : c \in C\}$
- Shattering: C shatters S if $|\Pi_C(S)| = 2^{|S|}$
- VC dimension: size of largest S shattered
  - $\max\{d: \exists S, |S| = d, \Pi_C(S) = 2^{|S|}\}$
  - If "no max" then infinity
    - For every d there is a shattered set of size d

VC dimension: Threshold

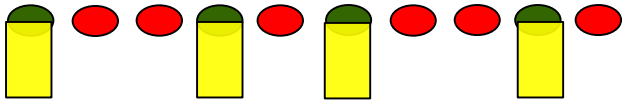
- $c_\theta(x) = I(x \geq \theta)$

- $VC \geq 1$
  - $S=\{0.5\}: c_{0.3}(0.5) = 1 \; and \; c_{0.6}(0.5) = 0$
- $VC < 2$
  - $S = \{z_1, z_2\}$
  - Assume $z_1 < z_2$
  - $c(z_1) = 1, c(z_2) = 0$

VC dimension: union of intervals

- Intervals on [0,1]

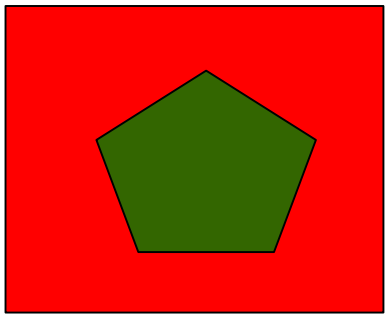
  - Finite but unbounded
- For any *d* points:



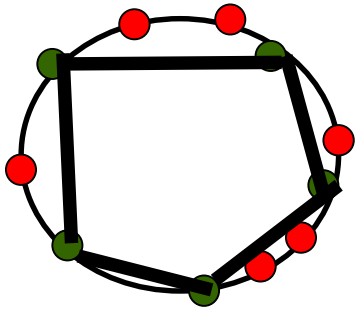
- VC-dim = infinity

VC dimension: convex polygon

- Convex polygon



- For any d points



- VC dimension = infinity

VC dimension: hyperplane

- $c_{w,\theta}(x) = sign(\sum_i w_i x_i + \theta)$
- VC dimension $\geq d+1$
  - $S = \{\vec{0}, \overrightarrow{e_1}, \ldots, \overrightarrow{e_d}\}$
  - Given labeling L in $\{-1,+1\}$, define $c_{w,\theta}(x)$:
    - $w_i = L(x_i)$
    - $\theta = \frac{L(0)}{2}$
  - $c_{w,\theta}(0) = sign(\theta)$ & $c_{w,\theta}(e_i) = sign(w_i + \theta)$

VC dimension: hyperplane

- VC dimension < $d+2$
  - For contradiction
    - Assume there is a shattered S and $|S|=d+2$
- Radom Theorem: $S \subseteq \mathbb{R}^d, |S| \geq d+2$
  - $\exists S' \subset S: \ conv(S') \cap conv(S - S') \neq \emptyset$
- Let S' be positive and S-S' be negative
  - Let $c_{w,\theta}$ be the separating hyperplane
    - Let POS be the positive and NEG be negatives of $c_{w,\theta}$

VC dimension: hyperplane

- $conv(S') \subset POS$ & $conv(S - S') \subset NEG$
  - closed under convex combinations
- Radom Theorem:
  - $conv(S') \cap conv(S - S') \neq \emptyset$
- However: $POS \cap NEG = \emptyset$
- Contradiction!
  - There is no such set S
- *VC-dim < d+2*

❑QED

VC dim: Sample lower bound

- Theorem:

VC-dim(C)=d+1

  - $m \geq \frac{d}{16\epsilon}$

- Proof:
  - Let $\{z_0, z_1, \ldots, z_d\}$
  - $D(x)$

$$= \begin{cases} 1-8\epsilon & x = z_0 \\ \frac{8\epsilon}{d} & x = z_i \\ 0 & otherwize \end{cases}$$

- Target function:
  - $c_t(z_0) = 1;$
  - $c_t(z_i) = 0 \; or \; 1 \; (prob \frac{1}{2})$
- $RARE = \{z_1, \ldots, z_d\}$
  - Assume $|S \cap RARE| \leq \frac{d}{2}$
  - $|UNSEEN| \geq \frac{d}{2}$
  - $\Pr[error] \geq \frac{1}{2}\frac{8\epsilon}{d}|UNSEEN| \geq 2\epsilon$

VC dim: Sample lower bound

- $E[|S \cap RARE|] = 8\epsilon m \leq \frac{d}{2}$
- $\Pr\left[|S \cap RARE| \leq \frac{d}{2}\right] \geq \frac{1}{2}$
- With probability at least ½
  - Error at least 2ε
- QED

VC dim: sample upper bound

- Incorrect proof
- For sample S:
  - $C_{|S} = \Pi_C(S)$ is finite
- Use finite class bound:
  - $m \geq \frac{1}{\epsilon}\log\frac{|\Pi_C(S)|}{\delta}$
- Problem:
  - S defines $C_{|S} = \Pi_C(S)$
- Solution
  - Take $2m$ points
    - $S = S_1 \cup S_2$
    - The randomization in the split to S_1 and S_2
  - Benefit:
    - We have $\Pi_C(S)$

VC dim: sample upper bound

- Bad concepts
  - Bad = $\{h|\text{error}(h)>\varepsilon\}$
- Hitting set *S*:
  - For every *h* in *Bad*
  - Exists *x* in *S*
  - $c_t(x) \neq h(x)$
- Goal
  - Compute prob. of S being a hitting set

- Event A:
  - S_1 not hitting set
    - Exists h in Bad which is consistent
  - Pr[A] < ???
- Event B:
  - Exists *h* in Bad
    - *h* consistent with S_1
    - *h* has ϵm errors on S_2

VC dim: sample upper bound

- Pr[B]=Pr[B|A]Pr[A]
  - Since B implies A
- Pr[B|A]
  - Fix such an h
  - Expected errors $\geq \varepsilon m$
  - Probability at least ½
- Result:
  - 2 Pr[B] $\geq$ Pr[A]

- $F = \Pi_C(S_1 \cup S_2)$
- Fix h in F:
  - h consistent with S_1
  - h has errors $\geq \varepsilon m$ on S_2
    - l number of errors
- Compute the prob. over partitions
  - S_1 and S_2

VC dim: sample upper bound

- Number of total partitions:
  - $\binom{2m}{l}$
- Number of partitions which make h consistent on S_1
  - $\binom{m}{l}$

- Prob bound
  - $\frac{\binom{m}{l}}{\binom{2m}{l}} = \prod_{i=0}^{l-1} \frac{m-i}{2m-i} \leq \frac{1}{2^l}$
- Bounding probabilities:
  - Union bound over h in F
  - $Pr[B] \leq |F| 2^{-\varepsilon m}$
  - $Pr[A] \leq 2Pr[B] \leq 2|F| 2^{-\varepsilon m}$

VC dim: sample upper bound

- High confidence
  - $\delta \geq 2|F|2^{-\epsilon m}$
  - $m \geq \frac{1}{\epsilon}\log\frac{2|F|}{\delta}$
- Need to bound |F|
  - $F = \Pi_C(S_1 \cup S_2)$

- Sauer-Shelah Lemma:
  - *VC-dim(C)=d*
  - $|S|=2m$
  - $\Pi_C(S) \leq \sum_{i=0}^{d} \binom{2m}{i}$
- Bound
  - 2^m for m≤d
  - $2(2m)^d$ for m>d

VC dim: Sampling Theorem

- Sample bound
  - $m \geq \frac{1}{\epsilon} \log \frac{4(2m)^d}{\delta}$
  - $m \geq \frac{2}{\epsilon} + \frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{d}{\epsilon} \log 2m$
  - $m = O\left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{d}{\epsilon} \log \frac{d}{\epsilon}\right)$
- Realizable case
- Non-realizable
  - $m = O\left(\frac{1}{\epsilon^2} \log \frac{1}{\delta} + \frac{d}{\epsilon^2} \log \frac{d}{\epsilon}\right)$
- Proof methodology

Rademacher Complexity

- Motivation:
  - Tighter bounds; Dist. Dependent
- Notation:
  - $f \in \{-1, +1\}; f \in F$
  - $\Pr[\sigma_i = +1] = \Pr[\sigma_i = -1] = \frac{1}{2}$

Rademacher Complexity

- Definition (Radmacher Complexity):
  - S sample of size m
  - $R_S(F) = E_\sigma[\max_{f \in F} \sum_{i=1}^{m} \sigma_i f(x_i)]$
  - $R_D(F) = E_S[R_S(F)]$

Rademacher Complexity: expected overfitting

- Theorem (expected overfitting):
  - $E_S\left[\max_{f\in F}\frac{1}{m}\sum_{i=1}^{m} f(x_i) - E_D[f(x)]\right] \le 2R_D(F)$
- Proof:
  - Two sample trick, add S'
  - $= E_S\left[\max_{f\in F}\frac{1}{m}\sum_{i=1}^{m} f(x_i) - E_{S'}[\frac{1}{m}\sum_{i=1}^{m} f(x_i')]\right]$
  - $\le E_{S,S'}\left[\max_{f\in F}\frac{1}{m}\sum_{i=1}^{m} f(x_i) - f(x_i')\right]$

Rademacher Complexity: expected overfitting

- $= E_{S,S'}\left[\max_{f\in F}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(f(x_i)-f(x_i'))\right]$
- $\leq E_S\left[\max_{f\in F}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(x_i)\right]+E_{S'}\left[\max_{f\in F}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(x_i')\right]$

- $= 2R_D(F)$
- QED

Rademachar Theorem

- With probability 1-δ, for every $h \in H$:
  - $\epsilon(h) \leq \hat{\epsilon}(h) + R_D(H) + \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2m}}$

    $\leq \hat{\epsilon}(h) + R_S(H) + 3\sqrt{\frac{\ln(\frac{2}{\delta})}{2m}}$

Model selection - Outline

- Motivation
- Overfitting
- Structural Risk Minimization
- Hypothesis Validation

Motivation:

- Problems:
  - We have too few examples
  - We have a very rich hypothesis class
  - How can we find the best hypothesis?
- Alternatively:
  - Usually we choose the hypothesis class
  - How rich of a class we want?
- How should we go about doing it?

Overfitting

- Concept class: Intervals on a line
- Can classify any training set
- Zero training error:
  - Is this the only goal?!

Overfitting: Intervals

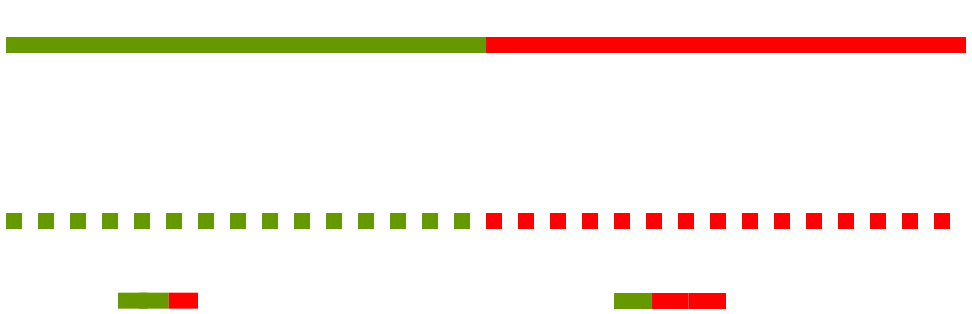
- Can always get zero training error!
- Are we interested in zero training error?!

Overfitting: Intervals

intervals	0	1	2	3	4
errors	7	3	2	1	0

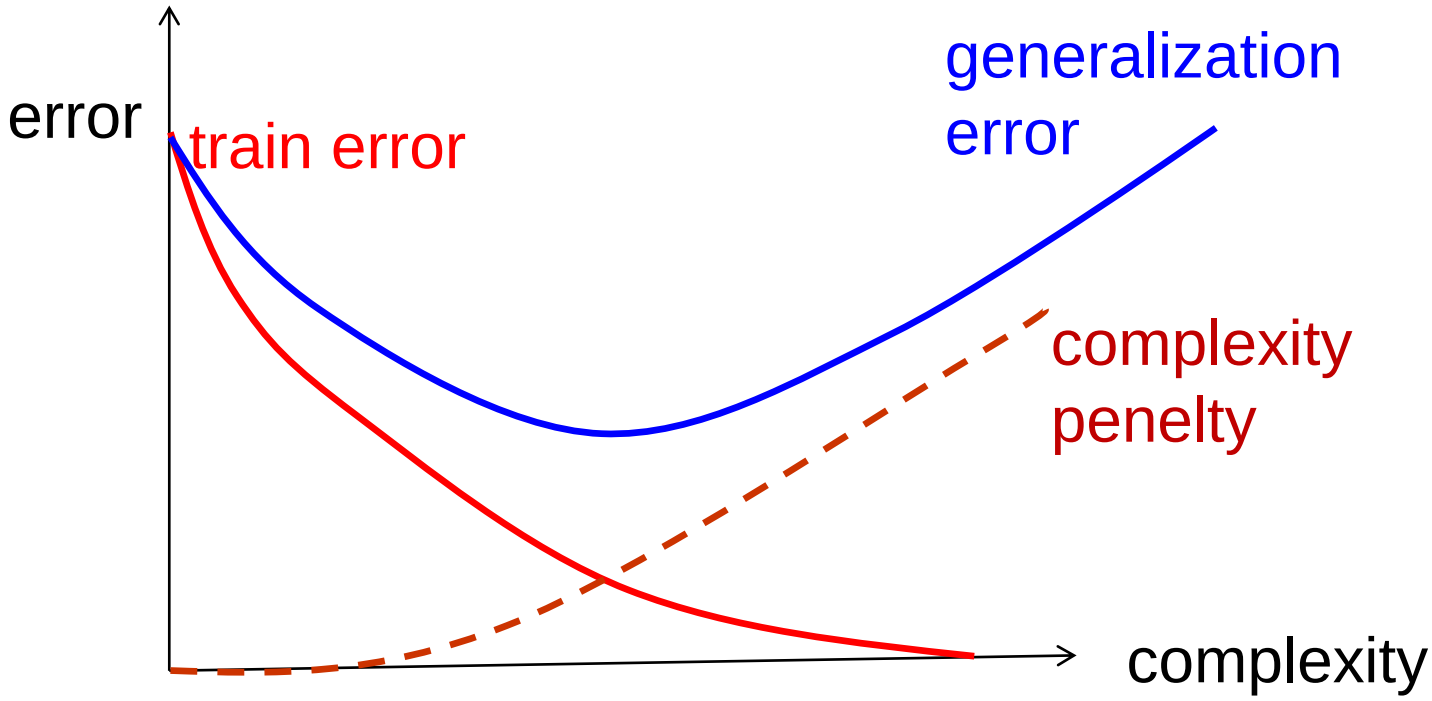
Overfitting

- Simple concept plus noise
- A very complex concept
  - insufficient number of examples



+ noise 1/3

Model Selection



Theoretical Model

- Nested Hypothesis classes
  - $H_1 \subseteq H_2 \subseteq H_3 \subseteq \ldots \subseteq H_i \subseteq$
- There is a target function $c_t(x)$,
  - non-realizable.
- True errors:
  - $\varepsilon(h) = Pr\,[\, h \neq c_t]$
  - $\varepsilon_i = \mathbf{inf}_{h \in Hi}\, e(h)$
  - $\varepsilon(h^*) = \mathbf{inf}_i\, \varepsilon_i$
    - *h^* is best hypothesis*
- Training error
  - $\hat{\epsilon}\,(h) = \frac{1}{m} \sum_{i=1}^{m} I[\, h \neq c_t]$
  - $\hat{\epsilon}_i = \mathbf{inf}_{h \in Hi}\, \hat{\epsilon}\,(h)$

Theoretical Model

- Complexity of h
  - $d(h) = \mathbf{min}_i \{h \in H_i\}$
- Add a penalty for $d(h)$
  - minimize: $\hat{\epsilon}(h)+penalty(h)$
- Penalty based.
  - Chose the hypothesis which minimizes:
  - $\hat{\epsilon}(h)+penalty(h)$

Structural Risk Minimization

- Parameters: λ_i and δ_i such that:

$$\Pr[\exists h \in H_i : |\hat{\epsilon}(h) - \epsilon(h)| > \lambda_i] \leq \delta_i$$

  - $\sum_i \delta_i = \delta$
    - $\delta_i = \delta/2^i$
- Implies: with prob. 1-δ
  - $\Pr[\exists h \in H : |\hat{\epsilon}(h) - \epsilon(h)| > \lambda_{d(h)}] \leq \delta_{d(h)}$

Structural Risk Minimization

- Setting $penalty(h) = \lambda_{d(h)}$
- Finite $|H_i|$
  - $\lambda_i = \sqrt{\frac{\log|H_i|/\delta}{m}}$
- $VC\text{-}dim(H_i)=i$
  - $\lambda_i = \sqrt{\frac{i \log\ i/\delta}{m}}$

SRM: Performance

- THEOROM
  - h^* : best hypothesis
  - g_{srm} : SRM choice
  - With probability $1-\delta$
  - $\varepsilon(h^*) \leq \varepsilon(g_{srm}) \leq \varepsilon(h^*) + 2\ penalty(h^*)$

- Note: bound depends only on h^*

Proof

- Bounding the error in H_i
  - $\Pr[|\hat{\epsilon}(g_{srm}) - \epsilon(g_{srm})| > \lambda_{srm}]$
    $\leq \Pr[\exists h \in H_{srm}: |\hat{\epsilon}(h) - \epsilon(h)| > \lambda_{srm}] \leq \delta_{srm}$
- Bounding the error across H_i
  - $\hat{\epsilon}(g_{srm}) \geq \epsilon(g_{srm}) - \lambda_{srm}$
  - $\hat{\epsilon}(h^*) + \lambda_* \geq \hat{\epsilon}(g_{srm}) + \lambda_{srm}$
  - $\epsilon(h^*) + \lambda_* \geq \hat{\epsilon}(h^*)$
- $\epsilon(h^*) + 2\lambda_* \geq \epsilon(g_{srm})$ QED

Hypothesis Validation

- Separate sample to training and selection.
- Using the training
  - Select from each H_i a candidate g_i
- Using the selection sample
  - select between $g_1, \ldots, g_m$
- The split size
  - $(1-\gamma)m$ training set
  - γm selection set

Hypothesis Validation: Algorithm

- Using (1-γ)m examples: S_1
  - $\widehat{\epsilon_1}(h)$= error on S_1
  - $g_i = \arg\min_{h \in H_i} \widehat{\epsilon_1}(h)$
- Using γm examples: S_2
  - $\widehat{\epsilon_2}(h)$= error on S_2
  - $g_{HV} = \arg\min_{g_i \in G} \widehat{\epsilon_2}(g_i)$
- Return g_{HV}

Hypo. Validation: Performance

- Errors
  - $\varepsilon_{hv}(m)$ = error of HV
    - Using m examples
  - $\varepsilon_{A}(m)$ = error of A
    - Any algorithm
    - Using m examples
    - Selecting g_i from H_i
      - only restriction on A
    - For example: any penalty function

- Theorem: with probability 1-δ

$$\varepsilon_{hv}(m) \le \varepsilon_{A}((1-\gamma)m) + 2\sqrt{\frac{\ln(2m/\delta)}{\gamma m}}$$

Hypo. Validation: Analysis

- $\Pr[|\epsilon(g_i) - \hat{\epsilon}_2(g_i)| > \lambda] \leq 2e^{-\lambda^2 \gamma m}$
- $\Pr[\exists i: |\epsilon(g_i) - \hat{\epsilon}_2(g_i)| > \lambda] \leq 2|G|e^{-\lambda^2 \gamma m} = \delta$
- Since $|G| \leq m$: $\lambda = \sqrt{\frac{\ln 2m/\delta}{\gamma m}}$
  - $\epsilon_2(g_i) + \lambda \geq \hat{\epsilon}_2(g_i)$
  - $\hat{\epsilon}_2(g_i) \geq \hat{\epsilon}_2(g_{HV})$
  - $\hat{\epsilon}_2(g_{HV}) \geq \epsilon(g_{HV}) - \lambda$
- $\epsilon_2(g_i) + 2\lambda \geq \epsilon(g_{HV})$

Summary

- PAC model
- Generalization bounds
  - Empirical Risk Minimization
  - VC dimension
  - Rademacher complexity
- Model Selection
  - Structural Risk Minimization (SRM)
  - Hypothesis selection